PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2002-264053
(43)Date of publication of application : 18.09.2002

(51)Int.CI.

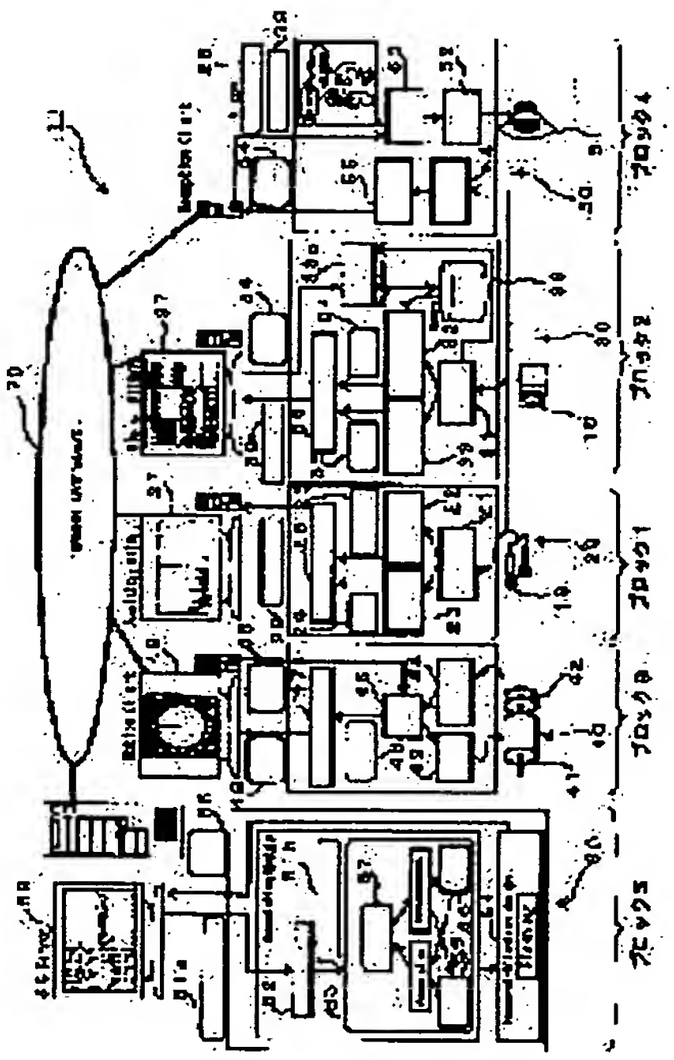
B25J 9/10
B25J 5/00
B25J 19/02
G06T 1/00
G06T 7/00
G06T 7/60
G10L 11/04
G10L 13/00
G10L 15/28
G10L 17/00
G10L 15/00
G10L 15/22
G10L 15/20
G10L 21/02
G10L 15/02
H04N 7/18

(21)Application number : 2001-067849
(22)Date of filing : 09.03.2001

(71)Applicant : JAPAN SCIENCE & TECHNOLOGY CORP
(72)Inventor : NAKADAI KAZUHIRO
HIDAI KENICHI
OKUNO HIROSHI
KITANO HIROAKI

(54) ROBOT AUDIO-VISUAL SYSTEM

(57)Abstract:
PROBLEM TO BE SOLVED: To provide a robot audio-visual system performing the visual and auditory tracing of an object to perform the audio-visual servo of a robot, using both visual and auditory senses.
SOLUTION: An auditory module 20 extracts an auditory event 28 by identifying the sound source of a speaker by pitch extraction and the separation and orientation of the sound source form an acoustic signal of a microphone. A visual module 30 extracts a visual event 39 by the face identification and orientation of the speaker from the image of a camera. A motor control module 40 extracts a motor event 49 from the rotating position of a drive motor. An association module 60 creates an auditory stream 65 and a visual stream 66 from the auditory event, visual event and motor event and creates an association stream 67 by associating these streams. An attention module 64 performs attention control on the basis of the association stream and performs the audio-visual servo of the robot.



LEGAL STATUS

[Date of request for examination] 11.07.2003
[Date of sending the examiner's decision of rejection]
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12)**公開特許公報**（Ａ）　(11)特許出願公開番号
特開2002－264053
(P2002－264053A)
(43)公開日　平成14年９月18日(2002.9.18)

(51)Int.Cl.'		識別記号	FI		テーマコード(参考)	
Ｂ２５Ｊ	9/10		Ｂ２５Ｊ	9/10	Ａ	３Ｃ００７
	5/00			5/00	Ｃ	５Ｂ０５７
	19/02			19/02		５Ｃ０５４
Ｇ０６Ｔ	1/00	３４０	Ｇ０６Ｔ	1/00	３４０Ａ	５Ｄ０１５
	7/00	３００		7/00	３００Ｆ	５Ｄ０４５

審査請求　未請求　請求項の数４　ＯＬ　（全 17 頁）　最終頁に続く

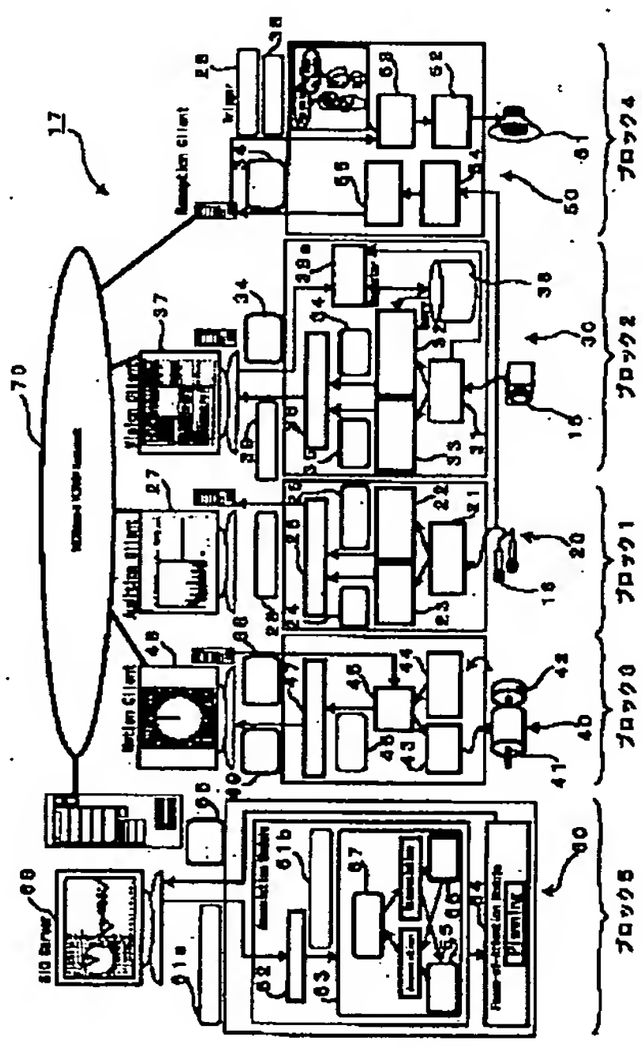
(21)出願番号　特願2001－67849(P2001－67849)
(22)出願日　平成13年３月９日(2001.3.9)

(71)出願人　396020800
科学技術振興事業団
埼玉県川口市本町４丁目１番８号
(72)発明者　中臺　一博
千葉県佐倉市臼井86
(72)発明者　日台　健一
埼玉県上尾市上野299－１
(72)発明者　奥乃　博
東京都渋谷区西原２－10－９
(72)発明者　北野　宏明
埼玉県川越市西小仙波町２－18－３
(74)代理人　100082876
弁理士　平山　一幸　（外１名）

最終頁に続く

(54)【発明の名称】　ロボット視聴覚システム

(57)【要約】
【課題】　対象に対する視覚及び聴覚の追跡を行なって、視覚及び聴覚の双方を使用してロボットの視聴覚サーボを行なうようにした、ロボット視聴覚システムを提供する。
【解決手段】　聴覚モジュール２０がマイクの音響信号からピッチ抽出，音源の分離及び定位により話者の音源を同定して聴覚イベント２８を抽出し、視覚モジュール３０がカメラの画像から話者の顔識別と定位により視覚イベント３９を抽出し、モータ制御モジュール４０が駆動モータの回転位置からモータイベント４９を抽出し、アソシエーションモジュール６０が聴覚イベント，視覚イベント及びモータイベントから聴覚ストリーム６５及び視覚ストリーム６６を生成し、これらを関連付けてアソシエーションストリーム６７を生成して、アテンション制御モジュール６４が、アソシエーションストリームに基づいてアテンション制御を行なって、ロボットの視聴覚サーボを行なう。



1

【特許請求の範囲】
【請求項１】　外部の音を集音する少なくとも一対のマイクを含む聴覚モジュールと、
ロボットの前方を撮像するカメラを含む視覚モジュールと、
ロボットを水平方向に回動させる駆動モータを含むモータ制御モジュールと、
前記聴覚モジュール，視覚モジュール及びモータ制御モジュールからのイベントを統合してストリームを生成するアソシエーションモジュールと、
アソシエーションモジュールにより生成されたストリームに基づいてアテンション制御を行なうアテンション制御モジュールと、を備えているロボット視聴覚システムであって、
前記聴覚モジュールが、マイクからの音響信号に基づいて、ピッチ抽出，音源の分離及び定位から、少なくとも一人の話者の音源を同定してその聴覚イベントを抽出し、
前記視覚モジュールが、カメラにより撮像された画像に基づいて、各話者の顔識別と定位からその視覚イベントを抽出し、
前記モータ制御モジュールが、駆動モータの回転位置に基づいて、モータイベントを抽出することにより、
前記アソシエーションモジュールが、聴覚イベント，視覚イベント及びモータイベントから、聴覚ストリーム及び視覚ストリームと、これらを関連付けたアソシエーションストリームを生成して、
前記アテンション制御モジュールが、アソシエーションストリームに基づいてモータ制御モジュールの駆動モータ制御のプランニングのためのアテンション制御を行なって、ロボットの視聴覚サーボを行なうことを特徴とする、ロボット視聴覚システム。
【請求項２】　外部の音を集音する少なくとも一対のマイクを含む聴覚モジュールと、
ロボットの前方を撮像するカメラを含む視覚モジュールと、
ロボットを水平方向に回動させる駆動モータを含むモータ制御モジュールと、
前記聴覚モジュール，視覚モジュール及びモータ制御モジュールからのイベントを統合してストリームを生成するアソシエーションモジュールと、
アソシエーションモジュールにより生成されたストリームに基づいてアテンション制御を行なうアテンション制御モジュールと、を備えている人型または動物型のロボットの視聴覚システムであって、
前記聴覚モジュールが、マイクからの音響信号に基づいて、ピッチ抽出，音源の分離及び定位から、少なくとも一人の話者の音源を同定してその聴覚イベントを抽出し、
前記視覚モジュールが、カメラにより撮像された画像に

2

基づいて、各話者の顔識別と定位からその視覚イベントを抽出し、
前記モータ制御モジュールが、駆動モータの回転位置に基づいてモータイベントを抽出することにより、
前記アソシエーションモジュールが、聴覚イベント，視覚イベント及びモータイベントから、聴覚ストリーム及び視覚ストリームと、これらを関連付けたアソシエーションストリームを生成して、
前記アテンション制御モジュールが、アソシエーションストリームに基づいてモータ制御モジュールの駆動モータ制御のプランニングのためのアテンション制御を行なって、ロボットの視聴覚サーボを行なうことを特徴とする、ロボット視聴覚システム。
【請求項３】　前記アテンション制御モジュールが、アテンション制御を行なう際に、アソシエーションストリーム，聴覚ストリーム及び視覚ストリームの順に優先させることを特徴とする、請求項１又は２に記載のロボット視聴覚システム。
【請求項４】　前記アテンション制御モジュールが、複数の聴覚ストリーム又は視覚ストリームが存在するとき、状況に応じて一つの聴覚ストリームまたは視覚ストリームを選択し、必要に応じてアソシエーションストリームを生成し、これらの聴覚ストリーム，視覚ストリーム又はアソシエーションストリームに基づいてアテンション制御を行なうことを特徴とする、請求項１から３の何れかに記載のロボット視聴覚システム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明はロボット、特に人型または動物型ロボットにおける視聴覚システムに関するものである。
【０００２】
【従来の技術】近年、このような人型または動物型ロボットにおいては、視覚，聴覚の能動知覚が注目されてきている。能動知覚とは、ロボット視覚やロボット聴覚等の知覚を担当する知覚装置を、知覚すべき対象に追従するように、これらの知覚装置を支持する例えば頭部を駆動機構により姿勢制御するものである。
【０００３】ここで、能動視覚に関しては、少なくとも知覚装置であるカメラが、駆動機構による姿勢制御によってその光軸方向が対象に向かって保持され、対象に対して自動的にフォーカシングやズームイン，ズームアウト等が行なわれることにより対象がカメラによって撮像されるようになっており、種々の研究が行なわれている。
【０００４】これに対して、能動聴覚に関しては、少なくとも知覚装置であるマイクが駆動機構による姿勢制御によって、その指向性が対象に向かって保持され、対象からの音がマイクによって集音される。このような能動聴覚は、例えば本出願人による特願２０００－２２６７

７号（ロボット聴覚システム）に開示されており、視覚情報を参照して音源の方向付けを行なうようにしている。

【０００５】

【発明が解決しようとする課題】ところで、これらの能動視覚及び能動聴覚は、ロボットの向き（水平方向）を変更するためのモータ制御モジュールと密接に関連があり、特定の対象に対して能動視覚及び能動聴覚を働かせるためには、ロボットを特定の対象に向ける、即ちアテンション制御を行なう必要がある。しかしながら、従来、所謂視覚サーボまたは聴覚サーボによるモータモジュールの駆動モータのアテンション制御は行なわれているが、視覚及び聴覚の双方を使用してロボットを正確に制御する、視聴覚サーボは行なわれていない。

【０００６】この発明は、以上の点にかんがみて、対象に対する視覚及び聴覚の追跡を行なって、視覚及び聴覚の双方を使用してロボットの視聴覚サーボを行なうようにした、ロボット視聴覚システムを提供することを目的としている。

【０００７】

【課題を解決するための手段】前記目的は、この発明によれば、外部の音を集音する少なくとも一対のマイクを含む聴覚モジュールと、ロボットの前方を撮像するカメラを含む視覚モジュールと、ロボットを水平方向に回動させる駆動モータを含むモータ制御モジュールと、聴覚モジュール，視覚モジュール及びモータ制御モジュールからのイベントを統合してストリームを生成するアソシエーションモジュールと、アソシエーションモジュールにより生成されたストリームに基づいてアテンション制御を行なうアテンション制御モジュールと、を備えているロボット視聴覚システムであって、聴覚モジュールが、マイクからの音響信号に基づいて、ピッチ抽出，音源の分離及び定位から、少なくとも一人の話者の音源を同定してその聴覚イベントを抽出し、視覚モジュールが、カメラにより撮像された画像に基づいて、各話者の顔識別と定位からその視覚イベントを抽出し、モータ制御モジュールが、駆動モータの回転位置に基づいて、モータイベントを抽出することにより、アソシエーションモジュールが、聴覚イベント，視覚イベント及びモータイベントから、聴覚ストリーム及び視覚ストリームと、これらを関連付けたアソシエーションストリームを生成して、アテンション制御モジュールが、アソシエーションストリームに基づいてモータ制御モジュールの駆動モータ制御のプランニングのためのアテンション制御を行なって、ロボットの視聴覚サーボを行なうことを特徴とするロボット視聴覚システムにより、達成される。

【０００８】また、前記目的は、この発明によれば、外部の音を集音する少なくとも一対のマイクを含む聴覚モジュールと、ロボットの前方を撮像するカメラを含む視覚モジュールと、ロボットを水平方向に回動させる駆動

モータを含むモータ制御モジュールと、聴覚モジュール，視覚モジュール及びモータ制御モジュールからのイベントを統合してストリームを生成するアソシエーションモジュールと、アソシエーションモジュールにより生成されたストリームに基づいてアテンション制御を行なうアテンション制御モジュールと、を備えている人型または動物型のロボットの視聴覚システムであって、聴覚モジュールが、マイクからの音響信号に基づいて、ピッチ抽出，音源の分離及び定位から少なくとも一人の話者の音源を同定してその聴覚イベントを抽出し、視覚モジュールが、カメラにより撮像された画像に基づいて、各話者の顔識別と定位からその視覚イベントを抽出し、モータ制御モジュールが、駆動モータの回転位置に基づいて、モータイベントを抽出することにより、アソシエーションモジュールが、聴覚イベント，視覚イベント及びモータイベントから、聴覚ストリーム及び視覚ストリームと、これらを関連付けたアソシエーションストリームを生成して、アテンション制御モジュールが、アソシエーションストリームに基づいてモータ制御モジュールの駆動モータ制御のプランニングのためのアテンション制御を行なって、ロボットの視聴覚サーボを行なうことを特徴とするロボット視聴覚システムにより、達成される。

【０００９】本発明によるロボット視聴覚システムは、好ましくは、前記アテンション制御モジュールが、アテンション制御を行なう際に、アソシエーションストリーム，聴覚ストリーム及び視覚ストリームの順に優先させる。

【００１０】本発明によるロボット視聴覚システムは、好ましくは、前記アテンション制御モジュールが、複数の聴覚ストリーム又は視覚ストリームが存在するとき、状況に応じて一つの聴覚ストリームまたは視覚ストリームを選択し、必要に応じてアソシエーションストリームを生成し、これらの聴覚ストリーム，視覚ストリームまたはアソシエーションストリームに基づいてアテンション制御を行なう。

【００１１】前記構成によれば、聴覚モジュールが、マイクが集音した外部の対象からの音から調波構造を利用してピッチ抽出を行なうことにより音源毎の方向を得て、個々の話者の音源を同定し、その聴覚イベントを抽出する。また、視覚モジュールが、カメラにより撮像された画像から、パターン認識による各話者の顔識別と定位から個々の話者の視覚イベントを抽出する。さらに、モータ制御モジュールが、ロボットを水平方向に回動させる駆動モータの回転位置に基づいて、ロボットの方向を検出することによってモータイベントを抽出する。なお、前記イベントとは、各時点において検出される音または顔が検出され、ピッチ及び方向等の特徴が抽出され、話者同定や顔識別等が行なわれること、あるいは駆動モータが回転される状態を示しており、ストリームと

は、時間的に連続するイベントを示している。

【００１２】ここで、アソシエーションモジュールは、このようにしてそれぞれ抽出された聴覚イベント，視覚イベント及びモータイベントに基づいて、各話者の聴覚ストリーム及び視覚ストリームを生成し、さらにこれらのストリームを関連付けてアソシエーションストリームを生成して、前記アテンション制御モジュールが、アソシエーションストリームに基づいてアテンション制御を行なうことにより、モータ制御モジュールの駆動モータ制御のプランニングを行なう。アテンションとは、ロボットが対象である話者を、聴覚的及び／又は視覚的に「注目」することであり、アンテンション制御とは、モータ制御モジュールによりその向きを変えることにより、ロボットが前記話者に注目するようにすることである。そして、アテンション制御モジュールは、このプランニングに基づいて、モータ制御モジュールの駆動モータを制御することにより、視聴覚サーボによってロボットの方向を対象である話者に向ける。これにより、ロボットが対象である話者に対して正対することにより、聴覚モジュールが当該話者の声を感度の高い正面方向にてマイクにより正確に集音，定位することができる共に、視覚モジュールが当該話者の画像をカメラにより良好に撮像することができるようになる。

【００１３】ここで、前記アテンション制御モジュールが、アソシエーションストリームに基づいてアテンション制御を行なうことにより、聴覚情報及び視覚情報の双方を使用して、ロボットの視聴覚サーボを行なうことにより、同一物体からの音声と顔（画像）が同一人に由来していることに基づいて、アテンション制御を行なうことができるので、何れかの情報、即ち聴覚情報又は視覚情報の何れかによる聴覚サーボ又は視覚サーボの場合と比較して、より正確にロボットのサーボを行うことができる。

【００１４】従って、このような聴覚モジュール，視覚モジュール及びモータ制御モジュールと、アソシエーションモジュール及びアテンション制御モジュールとの連携によって、ロボットの視聴覚サーボを行なうことにより、ロボットの聴覚及び視覚がそれぞれ有する曖昧性が互いに補完されることになり、所謂ロバスト性が向上し、複数の話者であっても、各話者をそれぞれ知覚することができる。

【００１５】

【発明の実施の形態】以下、図面に示した実施形態に基づいて、この発明を詳細に説明する。図１乃至図４はこの発明によるロボット視聴覚システムの一実施形態を備えた実験用の人型ロボットの全体構成を示している。図１において、人型ロボット１０は、４ＤＯＦ（自由度）のロボットとして構成されており、ベース１１と、ベース１１上にて一軸（垂直軸）周りに回動可能に支持された胴体部１２と、胴体部１２上にて、三軸方向（垂直軸，左右方向の水平軸及び前後方向の水平軸）の周りに揺動可能に支持された頭部１３と、を含んでいる。

【００１６】前記ベース１１は固定配置されていてもよく、脚部として動作可能としてもよい。また、ベース１１は移動可能な台車等の上に載置されていてもよい。前記胴体部１２は、ベース１１に対して垂直軸の周りに、図１にて矢印Ａで示すように回動可能に支持されており、図示しない駆動手段によって回転駆動されると共に、図示の場合、防音性の外装によって覆われている。

【００１７】前記頭部１３は胴体部１２に対して連結部材１３ａを介して支持されており、この連結部材１３ａに対して前後方向の水平軸の周りに、図１にて矢印Ｂで示すように揺動可能に、また左右方向の水平軸の周りに、図２にて矢印Ｃで示すように揺動可能に支持されていると共に、前記連結部材１３ａが、胴体部１２に対してさらに前後方向の水平軸の周りに、図１にて矢印Ｄで示すように揺動可能に支持されており、それぞれ図示しない駆動手段によって、各矢印Ａ，Ｂ，Ｃ，Ｄ方向に回転駆動される。

【００１８】ここで、前記頭部１３は、図３に示すように全体が防音性の外装１４により覆われていると共に、前側にロボット視覚を担当する視覚装置としてのカメラ１５を、また両側にロボット聴覚を担当する聴覚装置としての一対のマイク１６（１６ａ，１６ｂ）を備えている。

【００１９】前記外装１４は、例えばウレタン樹脂等の吸音性の合成樹脂から構成されており、頭部１３の内部をほぼ完全に密閉することにより、頭部１３の内部の遮音を行なうように構成されている。尚、胴体部１２の外装も、同様にして吸音性の合成樹脂から構成されている。前記カメラ１５は公知の構成であって、例えば所謂パン，チルト，ズームの３ＤＯＦ（自由度）を有する市販のカメラが適用され得る。

【００２０】前記マイク１６は、それぞれ頭部１３の側面において、前方に向かって指向性を有するように取り付けられている。ここで、マイク１６の左右の各マイク１６ａ，１６ｂは、それぞれ図１及び図２に示すように、外装１４の両側にて前方に向いた段部１４ａ，１４ｂにて、内側に取り付けられ、段部１４ａ，１４ｂに設けられた貫通穴を通して、前方の音を集音すると共に、外装１４の内部の音を拾わないように適宜の手段により遮音されている。これにより、マイク１６ａ，１６ｂは、所謂バイノーラルマイクとして構成されている。なお、マイク１６ａ，１６ｂの取付位置の近傍において、外装１４は人間の外耳形状に形成されていてもよい。

【００２１】図４は、前記マイク１６及びカメラ１５を含むロボット視聴覚システムの電気的構成を示している。図４において、視聴覚システム１７は、パーティ受付及びコンパニオン用ロボットとしての構成であり、聴覚モジュール２０，視覚モジュール３０，モータ制御モ

ジュール４０，対話モジュール５０，アソシエーションモジュール６０及びアテンション制御モジュール６４と、から構成されている。以下、図４の各部を拡大して示す図５～図９をも参照しつつさらに説明する。説明の便宜上、聴覚モジュール２０をブロック１として図５に拡大して示し、視覚モジュール３０をブロック２として図６に拡大して示し、モータ制御モジュール４０をブロック３として図７に拡大して示し、対話モジュール５０をブロック４として図８に拡大して示し、さらに、アソシエーションモジュール６０及びアテンション制御モジュール６４をブロック５として図９に拡大して示す。ここで、アソシエーションモジュール６０（ブロック５、図９）はサーバから構成されていると共に、他のモジュール、即ち聴覚モジュール２０（ブロック１、図５），視覚モジュール３０（ブロック２、図６），モータ制御モジュール４０（ブロック３、図７），対話モジュール５０（ブロック４、図８）は、それぞれクライアントから構成されており、互いに非同期で動作する。

【００２２】なお、前記サーバ及び各クライアントは、例えばパーソナルコンピュータにより構成されており、例えば１００Ｂａｓｅ－Ｔ等のネットワーク７０を介して、例えばＴＣＰ／ＩＰプロトコルにより、相互にＬＡＮ接続されている。また、各モジュール２０，３０，４０，５０，６０は、それぞれ階層的に分散して、具体的には下位から順次にデバイス層，プロセス層，特徴層，イベント層から構成されている。

【００２３】図５に示すように、前記聴覚モジュール２０は、デバイス層としてのマイク１６と、プロセス層としてのピーク抽出部２１，音源定位部２２，音源分離部２３と、特徴層（データ）としてのピッチ２４，水平方向２５と、イベント層としての聴覚イベント生成部２６及びビューア２７と、から構成されている。

【００２４】これにより、聴覚モジュール２０は、マイク１６からの音響信号に基づいて、ピーク抽出部２１により左右のチャンネル毎に一連のピークを抽出して、左右のチャンネルで同じか類似のピークをペアとする。ここで、ピーク抽出は、パワーがしきい値以上で且つ極大値であって、例えば９０Ｈｚ乃至３ｋＨｚの間の周波数であるという条件のデータのみを透過させる帯域フィルタを使用することにより行なわれる。このしきい値は、周囲の暗騒音を計測して自動的に決定される。

【００２５】そして、聴覚モジュール２０は、各ピーク＊が調波構造を有していることを利用して、左右のチャンネル間でより正確なピークのペアを見つけ、左右のチャンネルのピークの各ペアについて、音源分離部２３により、逆ＦＦＴ（高速フーリエ変換）を適用して、各音源からの混合音から調波構造を有する音を分離する。これにより、聴覚モジュール２０は、分離した各音について音源定位部２２により左右のチャンネルから同じ周波数の音響信号を選択して、例えば５度毎にＩＰＤ（両耳間位相差）及びＩＩＤ（両耳間強度差）を求める。

【００２６】そして、聴覚モジュール２０の音源定位部２２は、所謂聴覚エピポーラ幾何を利用して、ロボット１０の正面を０度として±９０度の範囲で、仮説推論によるＩＰＤ　Ｐｈの仮説を生成して、

【数１】

$$d(\theta)=\frac{1}{n_{f<1.5\mathrm{KHz}}}\sum_{f=P0}^{1.5\mathrm{KHz}}\frac{(P_h(\theta,f)-P_s(f))^2}{f}$$

により分離した音と各仮説間の距離ｄ（θ）を計算する。ここで、ｎｆ＜１．５ｋＨｚは、周波数が１．５ｋＨｚ以下である倍音である。これは、左右のマイク１５のベースラインからＩＰＤが１．２乃至１．５ｋＨｚ以下の周波数に対して有効であるので、今回の実験では１．５ｋＨｚ以下としたものである。

【００２７】ＩＩＤについては、ＩＰＤと同様に、分離した音の各倍音の左右チャンネル間のパワー差から求められる。ただし、ＩＩＤについては仮説推論ではなく、

【数２】

$$I=\sum_{f=1.5\mathrm{KHz}}^{3\mathrm{KHz}}I_s(f)$$

による判別関数を用いて、音源が左右何れかを判定するものとする。即ち、周波数ｆの各倍音のＩＩＤをＩｓ（ｆ）としたとき、音源は、Ｉが正であればロボットの左方向に、Ｉが負であれば右方向に、そしてほぼ０であれば正面方向に存在することになる。ここで、ＩＩＤの仮説生成には、ロボット１０の頭部形状を考慮した膨大な計算が必要となることから、リアルタイム処理を考慮して、ＩＰＤと同様の仮説推論は行なわない。

【００２８】そして、聴覚モジュール２０の音源定位部２２は、前記距離ｄ（θ）から確立密度関数

【数３】

$$BF_{\mathrm{IPD}}(\theta)=\int_{-\infty}^{\frac{d(\theta)-m}{\sqrt{\frac{s}{n}}}}\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)dx$$

を利用して、ＩＰＤの確信度ＢＦ$_{IPD}$（θ）を計算する。ここで、ｍ，ｓは、それぞれｄ（θ）の平均と分散であり、ｎはｄの個数である。また、ＩＩＤの確信度ＢＦ$_{IID}$（θ）は、３０度＜θ≦９０度で、前記Ｉが＋のとき０．３５、－のとき０．６５、－３０度＜θ≦３０度で、前記Ｉが＋のとき０．５、－のとき０．５、－９０度＜θ≦－３０度で、前記Ｉが＋のとき０．６５、－のとき０．３５となる。

【００２９】そして、このようにして得られたＩＰＤの確信度ＢＦ$_{IPD}$（θ）及びＩＩＤの確信度ＢＦ$_{IID}$（θ）を、

【数４】

$$BF_{IPD+IID}(\theta) = BF_{IPD}(\theta)BF_{IID}(\theta) + (1 - BF_{IPD}(\theta))BF_{IID}(\theta) + BF_{IPD}(\theta)(1 - BF_{IID}(\theta))$$

で示されるＤｅｍｐｓｔｅｒ－Ｓｈａｆｅｒ理論により統合して、確信度ＢＦ$_{IPD+IID}$（θ）を生成する。これにより、聴覚モジュール２０は、聴覚イベント生成部２６により、音源方向として尤度の高い順に上位２０個の確信度ＢＦ$_{IPD+IID}$（θ）と方向（θ）のリストと、ピッチにより、聴覚イベント２８を生成する。

【００３０】このようにして、聴覚モジュール２０は、マイク１６からの音響信号に基づいて、ピッチ抽出，音源の分離及び定位から、少なくとも一人の話者の音源を同定してその聴覚イベントを抽出し、ネットワーク７０を介してアソシエーションモジュール６０に対して送信するようになっている。尚、聴覚モジュール２０における上述した処理は、４０ｍ秒毎に行なわれる。

【００３１】ビューア２７は、このようにして生成された聴覚イベント２８をクライアントの画面上に表示するものであり、具体的には図１１（Ａ）に示すように、左側のウインドウ２７ａに、聴覚イベント２８のパワースペクトルと抽出したピークを、右側のウインドウ２７ｂに、縦軸を相対的な方位角，横軸をピッチ（周波数）とする聴覚イベント２８のグラフを表示する。ここで、聴覚イベントは、音源定位の確信度を円の直径とする円により表現されている。

【００３２】図６に示すように、前記視覚モジュール３０は、デバイス層としてのカメラ１５と、プロセス層としての顔発見部３１，顔識別部３２，顔定位部３３と、特徴層（データ）としての顔ＩＤ３４，顔方向３５と、イベント層としての視覚イベント生成部３６及びビューア３７と、から構成されている。

【００３３】これにより、視覚モジュール３０は、カメラからの画像信号に基づいて、顔発見部３１により例えば肌色抽出により各話者の顔を検出し、顔識別部３２にて前もって登録されている顔データベース３８により検索して、一致した顔があった場合、その顔ＩＤ３４を決定して当該顔を識別すると共に、顔定位部３３により当該顔方向３５を決定（定位）する。なお、顔識別部３２による顔データベース３８の検索の結果、一致した顔がなかった場合には、顔学習部３８ａが、顔発見部３１が検出した顔を顔データベース３８に登録する。

【００３４】ここで、視覚モジュール３０は、顔発見部３１が画像信号から複数の顔を見つけた場合、各顔について前記処理、即ち識別及び定位そして追跡を行なう。その際、顔発見部３１により検出された顔の大きさ，方向及び明るさがしばしば変化するので、顔発見部３１は、顔領域検出を行なって、肌色抽出と相関演算に基づくパターンマッチングの組合せによって、２００ｍ秒以内に複数の顔を正確に検出できるようになっている。

【００３５】また、顔識別部３２は、顔発見部３１により検出された各顔領域画像を、判別空間に射影し、顔データベース３８に前もって登録された顔データとの距離ｄを計算する。この距離ｄは、登録顔数（Ｌ）に依存するので，

【数５】

$$P_v = \int_{\frac{d^2}{2}}^{\infty} e^{-t} t^{\frac{L}{2}-1} dt$$

により、パラメータの依存しない確信度Ｐｖに変換される。ここで、判別空間の基底となる判別行列は、公知のオンラインＬＤＡにより，通常のＬＤＡと比較して少ない計算により更新され得るので，リアルタイムに顔データを登録することが可能である。

【００３６】顔定位部３３は、二次元の画像平面における顔位置を三次元空間に変換し、顔が画像平面にて（ｘ，ｙ）に位置する幅と高さがそれぞれＸ及びＹであるｗ×ｗピクセルであるとすると、三次元空間における顔位置は、以下の各式で与えられる方位角θ，高さφ及び距離ｒのセットとして得られる。

【数６】

$$r = \frac{C_1}{w}$$

【数７】

$$\theta = \sin^{-1}\left(\frac{x - \frac{X}{2}}{C_2 r}\right)$$

【数８】

$$\phi = \sin^{-1}\left(\frac{\frac{Y}{2} - y}{C_2 r}\right)$$

ここで、Ｃ$_1$及びＣ$_2$は、探索画像サイズ（Ｘ，Ｙ）とカメラの画角そして実際の顔の大きさにより定義される定数である。

【００３７】そして、視覚モジュール３０は、各顔毎に、顔ＩＤ（名前）３４及び顔方向３５から、視覚イベント生成部３６により視覚イベント３９を生成する。詳細には、視覚イベント３９は、各顔毎に、上位５つの確信度付きの顔ＩＤ（名前）３４と位置（距離ｒ，水平角度θ及び垂直角度φ）から構成されている。

【００３８】なお、ビューア３７は、視覚イベントをクライアントの画面上に表示するものであり、具体的には図１１（Ｂ）に示すように、カメラ１５による画像３７ａと、顔識別の確信度付きで抽出した顔の顔ＩＤと定位の結果である位置のリスト３７ｂを表示する。ここで、

カメラ１５による画像には、発見し同定された顔が長方形の枠３７ｃで囲まれて表示されている。複数の顔が発見された場合には、各顔について、それぞれ同定を示す長方形の枠３７ｃと、定位の結果としてのリスト３７ｂが表示される。

【００３９】図７に示すように、前記モータ制御モジュール４０は、デバイス層としてのモータ４１及びポテンショメータ４２と、プロセス層としてのＰＷＭ制御回路４３，ＡＤ変換回路４４及びモータ制御部４５と、特徴層としてのロボット方向４６と、イベント層としてのモータイベント生成部４７と、ビューア４８と、から構成されている。

【００４０】これにより、モータ制御モジュール４０は、アテンション制御モジュール６４（後述）からの指令に基づいてモータ制御部４５によりＰＷＭ制御回路４３を介してモータ４１を駆動制御すると共に、モータ４１の回転位置をポテンショメータ４２により検出して、ＡＤ変換回路４４を介してモータ制御部４５によりロボット方向４６を抽出し、モータイベント生成部４７によりモータ方向情報から成るモータイベント４９を生成する。

【００４１】ビューア４８は、モータイベントをクライアントの画面上に三次元的に表示するものであって、具体的には図１２（Ａ）に示すように、モータイベント４９によるロボット１０の向きと動作速度を、例えばＯｐｅｎＧＬにより実装されている三次元ビューアを利用してリアルタイムに三次元表示するようになっている。

【００４２】図８に示すように、前記対話モジュール５０は、デバイス層としてのスピーカ５１及びマイク１６と、プロセス層としての音声合成回路５２，対話制御回路５３及び自声抑制回路５４，音声認識回路５５と、から構成されている。

【００４３】これにより、対話モジュール５０は、後述するアソシエーションモジュール６０により対話制御回路５３を制御し、音声合成回路５２によりスピーカ５１を駆動することによって、対象とする話者に対して所定の音声を発すると共に、マイク１６からの音響信号から自声抑制回路５４によりスピーカ５１からの音を除去した後、音声認識回路５５により対象とする話者の音声を認識する。なお、前記対話モジュール５０は、階層としての特徴層及びイベント層を備えていない。

【００４４】ここで、対話制御回路５３は、例えばパーティ受付ロボットの場合には、現在のアテンションを継続することが最優先となるが、パーティロボットの場合には、最も最近に関連付けられたストリームに対して、アテンション制御される。

【００４５】図９に示すように、前記アソシエーションモジュール６０は、上述した聴覚モジュール２０，視覚モジュール３０，モータ制御モジュール４０，対話モジュール５０に対して、階層的に上位に位置付けられており、各モジュール２０，３０，４０，５０のイベント層の上位であるストリーム層を構成している。具体的には、前記アソシエーションモジュール６０は、聴覚モジュール２０，視覚モジュール３０及びモータ制御モジュール４０からの非同期イベント６１ａ即ち聴覚イベント２８，視覚イベント３９及びモータイベント４９を同期させて同期イベント６１ｂにする同期回路６２と、これらの同期イベント６１ｂを相互に関連付けて、聴覚ストリーム６５，視覚ストリーム６６及びアソシエーションストリーム６７を生成するストリーム生成部６３と、さらにアテンション制御モジュール６４と、ビューア６８を備えている。

【００４６】前記同期回路６２は、聴覚モジュール２０からの聴覚イベント２８，視覚モジュール３０からの視覚イベント３８及びモータ制御モジュール４０からのモータイベント４９を同期させて、同期聴覚イベント，同期視覚イベント及び同期モータイベントを生成する。その際、聴覚イベント２８及び視覚イベント３８は、同期モータイベントによって、その座標系が絶対座標系に変換されることになる。

【００４７】ここで、各イベントの実際に観測されてからネットワーク７０を介してアソシエーションモジュール６０に到着するまでの遅延時間は、例えば聴覚イベント２８では４０ｍ秒、視覚イベント３９では２００ｍ秒、モータイベント４９では１００ｍであり、ネットワーク７０における遅延が１０乃至２００ｍ秒であり、さらに到着周期も異なることによるものである。従って、各イベントの同期を取るために、聴覚モジュール２０，視覚モジュール３０及びモータ制御モジュール４０からの聴覚イベント２８，視覚イベント３９及びモータイベント４９は、それぞれ実際の観測時間を示すタイムスタンプ情報を備えており、図示しない短期記憶回路にて、例えば２秒間の間だけ一旦記憶される。

【００４８】そして、同期回路６２は、短期記憶回路に記憶された各イベントを、上述した遅延時間を考慮して、実際の観測時間と比較して５００ｍ秒の遅延時間を備えるように、同期プロセスにより取り出す。これにより、同期回路６２の応答時間は５００ｍ秒となる。また、このような同期プロセスは例えば１００ｍ秒周期で動作するようになっている。尚、各イベントは、それぞれ互いに非同期でアソシエーションモジュール６０に到着するので、同期を取るための観測時刻と同時刻のイベントが存在するとは限らない。従って、同期プロセスは、同期を取るための観測時刻前後に発生したイベントに対して、線形補間による補間を行なうようになっている。

【００４９】また、ストリーム生成部６３は、以下の点に基づいて、ストリーム６５，６６，６７の生成を行なう。

１．　聴覚イベント２８は、同等または倍音関係にある

ピッチを備え、方向が±１０度以内で最も近い聴覚ストリーム６５に接続される。なお、±１０度以内の値は、聴覚エピポーラ幾何の精度を考慮して選定されたものである。

２．　視覚イベント３９は、共通の顔ＩＤ３４を有し且つ４０ｃｍの範囲内で最も近い視覚ストリーム６６に接続される。なお、４０ｃｍの範囲内の値は、秒速４ｍ以上で人間が移動することがないということを前提として選定されたものである。

３．　すべてのストリームに対して探索を行なった結果、接続可能なストリーム６５，６６が存在しないイベントがある場合には、当該イベント２８，３９は、新たなストリーム６５，６６を構成することになる。

４．　既に存在しているストリーム６５，６６は、これらに接続されるイベント２８，３９がない場合には、最大で５００ｍ秒間は存続するが、その後もイベントが接続されない状態が継続すると、消滅する。

５．　聴覚ストリーム６５と視覚ストリーム６６が±１０度以内に近接する状態が１秒間のうち５００ｍ秒以上継続する場合、これの聴覚ストリーム６５と視覚ストリーム６６は、同一話者に由来するものであるとみなされ、互いに関係付けられて、アソシエーションストリーム６７が生成される。

６．　アソシエーションストリーム６７は、聴覚イベント２８または視覚イベント３９が３秒間以上接続されない場合には、関係付けが解除され、既存の聴覚ストリーム６５または視覚ストリーム６６のみが存続する。

７．　アソシエーションストリーム６７は、聴覚ストリーム６５及び視覚ストリーム６６の方向差が３秒間、±３０度以上になった場合には、関係付けが解除され、個々の聴覚ストリーム６５及び視覚ストリーム６６に戻る。

【００５０】これにより、ストリーム生成部６３は、同期回路６２からの同期聴覚イベント及び同期視覚イベントに基づいて、これらの時間的つながりを考慮してイベントを接続することにより、聴覚ストリーム６５及び視覚ストリーム６６を生成すると共に、相互の結び付きの強い聴覚ストリーム６５及び視覚ストリーム６６を関係付けて、アソシエーションストリーム６７を生成するようになっており、逆にアソシエーションストリーム６７を構成する聴覚ストリーム６５及び視覚ストリーム６６の結び付きが弱くなれば、関係付けを解除するようになっている。

【００５１】また、アテンション制御モジュール６４は、モータ制御モジュール４０の駆動モータ制御のプランニングのためのアテンション制御を行なうものであり、その際アソシエーションストリーム６７，聴覚ストリーム６５そして視覚ストリーム６６の順に優先的に参照して、アテンション制御を行なう。そして、アテンション制御モジュール６４は、聴覚ストリーム６５及び視

覚ストリーム６６の状態とアソシエーションストリーム６７の存否に基づいて、ロボット１０の動作プランニングを行ない、駆動モータ４１の動作の必要があれば、モータ制御モジュール４０に対して動作指令としてのモータイベントをネットワーク７０を介して送信する。

【００５２】ここで、アテンション制御モジュール６４におけるアテンション制御は、連続性とトリガに基づいており、連続性により同じ状態を保持しようとし、トリガにより最も興味のある対象を追跡しようとする。

従って、アテンション制御は、１．　アソシエーションストリームの存在は、ロボット１０に対して正対して話している人が現在も存在している、あるいは近い過去に存在していたことを示しているので、このようなロボット１０に対して話している人に対して、高い優先度でアテンションを向けて、トラッキングを行なう必要がある。

２．　マイク１６は無指向性であるので、カメラの視野角のような検出範囲が存在せず、広範囲の聴覚ストリームを得ることができるので、視覚ストリームより聴覚ストリームの優先度を高くすべきである。という二つの点を考慮して、以下の原則に従ってアテンションを向けるストリームを選択して、トラッキングを行なう。

１．　アソシエーションストリームのトラッキングを最優先する。

２．　アソシエーションストリームが存在しない場合、聴覚ストリームのトラッキングを優先する。

３．　アソシエーションストリーム及び聴覚ストリームが存在しない場合、視覚ストリームのトラッキングを優先する。

４．　複数の同じ種類のストリームが存在する場合、最も古いストリームのトラッキングを優先する。

このようにして、アテンション制御モジュール６４は、聴覚情報及び視覚情報に基づいて生成されたアソシエーションストリームによりアテンション制御を行なって、ロボットの視聴覚サーボによりモータ制御モジュール４０の駆動モータ４１の制御のプランニングを行ない、このプランニングに基づいてモータコマンド６６を生成し、ネットワーク７０を介してモータ制御モジュール４０に伝送する。これにより、モータ制御モジュール４０では、このモータコマンド６６に基づいてモータ制御部４５がＰＷＭ制御を行なって、駆動モータ４１を回転駆動させて、ロボット１０を所定方向に向けるようになっている。

【００５３】ビューア６８は、このようにして生成された各ストリームをサーバの画面上に表示するものであり、具体的には図１２（Ｂ）に示すように、レーダチャート６８ａ及びストリームチャート６８ｂにより表示する。ここで、レーダチャート６８ａは、その瞬間におけるストリームの状態、より詳細には広く明るい扇形で示されるカメラの視野角６８ａ１と、狭く暗い扇形で示さ

れる音源方向６８ａ２を示す。また、ストリームチャート６８ｂは、太線により示されるアソシエーションストリーム６８ｂ１と、細線により示される聴覚ストリームまたは視覚ストリーム６８ｂ２を示している。

【００５４】本発明実施形態による人型ロボット１０は以上のように構成されており、パーティ受付ロボットとして対象とする話者に対して、図１０を参照して、以下のように動作する。先ず、図１０（Ａ）に示すように、ロボット１０は、パーティ会場の入口前に配置されている。そして、図１０（Ｂ）に示すように、パーティ参加者Ｐがロボット１０に接近してくるが、ロボット１０は、まだ当該参加者Ｐを認識していない。ここで、参加者Ｐがロボット１０に対して例えば「こんにちは」と話し掛けると、ロボット１０は、マイク１６が当該参加者Ｐの音声を拾って、聴覚モジュール２０が音源方向を伴う聴覚イベント２８を生成して、ネットワーク７０を介してアソシエーションモジュール６０に伝送する。

【００５５】これにより、アソシエーションモジュール６０は、この聴覚イベント２８に基づいて聴覚ストリーム２９を生成する。このとき、視覚モジュール３０は参加者Ｐがカメラ１５の視野内に入っていないので、視覚イベント３９を生成しない。従って、アソシエーションモジュール６０は、聴覚イベント２８のみに基づいて聴覚ストリーム２９を生成し、アテンション制御モジュール６４は、この聴覚ストリーム２９をトリガーとして、ロボット１０を参加者Ｐの方向に向けるようなアテンション制御を行なう。

【００５６】このようにして、図１０（Ｃ）に示すように、ロボット１０が参加者Ｐの方向を向き、所謂声によるトラッキングが行なわれる。そして、視覚モジュール３０がカメラ１５による参加者Ｐの顔の画像を取り込んで、視覚イベント３９を生成して、当該参加者Ｐの顔を顔データベース３８により検索し、顔識別を行なうと共に、その結果である顔ＩＤ３４及び画像をネットワーク７０を介してアソシエーションモジュール６０に伝送する。なお、当該参加者Ｐの顔が顔データベース３８に登録されていない場合には、視覚モジュール３０は、その旨をネットワーク７０を介してアソシエーションモジュールに伝送する。

【００５７】このとき、ロボット１０は、聴覚イベント２８及び視覚イベント３９によりアソシエーションストリーム６５を生成しており、このアソシエーションストリーム６５により視聴覚サーボを行なうことにより、アテンション制御モジュール６４は、そのアテンション制御を変更しないので、ロボット１０は、参加者Ｐの方向を向き続ける。従って、参加者Ｐが移動したとしても、ロボット１０は、アソシエーションストリーム６５によりモータ制御モジュール４０を制御することにより参加者Ｐを追跡して、視覚モジュール３０のカメラ１５が参加者Ｐを継続して撮像し得るようになっている。

【００５８】そして、アソシエーションモジュール６０は、聴覚モジュール２０の音声認識回路５５に入力を与えて、音声認識回路５５はその音声認識結果を対話制御回路５３に与える。これにより、対話制御回路５３は、音声合成を行なってスピーカ５１から発声する。このとき、音声認識回路５５がマイク１６からの音響信号からスピーカ５１からの音を自声抑制回路５４により低減することにより、ロボット１０は自身の発声を無視して相手の声をより正確に認識することができる。

【００５９】ここで、音声合成による発声は、参加者Ｐの顔が前記顔データベース３８に登録されているか否かで異なる。参加者Ｐの顔が顔データベース３８に登録されている場合には、アソシエーションモジュール６０は、視覚モジュール３０からの顔ＩＤ３４に基づいて、対話モジュール５０を制御して、音声合成により「こんにちは。ＸＸＸさんですか？」と参加者Ｐに対して質問する。これに対して、参加者Ｐが「はい。」と答えると、対話モジュール５０がマイク１６からの音響信号に基づいて、音声認識回路５５により「はい」を認識して、対話制御回路５３により音声合成を行ない、スピーカ５１から「ようこそＸＸＸさん、どうぞ部屋にお入り下さい。」と発声する。

【００６０】また、参加者Ｐの顔が顔データベース３８に登録されていない場合には、アソシエーションモジュール６０は、対話モジュール５０を制御して、音声合成により「こんにちは。あなたのお名前を教えていただけますか？」と参加者Ｐに対して質問する。これに対して、参加者Ｐが「ＸＸＸです。」と自分の名前を答えると、対話モジュール５０がマイク１６からの音響信号に基づいて、音声認識回路５５により「ＸＸＸ」を認識して、対話制御回路５３により音声合成を行ない、スピーカ５１から「ようこそＸＸＸさん、どうぞ部屋にお入り下さい。」と発声する。このようにして、ロボット１０は、参加者Ｐの認識を行なって、図１０（Ｄ）に示すように、パーティ会場への入場を誘導すると共に、視覚モジュール３０にて、当該参加者Ｐの顔の画像と名前「ＸＸＸ」を顔データベース３８に登録させる。

【００６１】また、人型ロボット１０は、コンパニオン用ロボットとして、図１３及び図１４を参照して、以下のように動作する。先ず、人型ロボット１０は、特に明確なシナリオを有しているのではなく、例えば図１３に示すシナリオをベンチマークとして使用して、本システムの評価を行なった。なお、図１４（Ａ）はロボット方向，図１４（Ｂ）は視覚ストリームによるトラッキング，図１４（Ｃ）は聴覚ストリームによるトラッキングを示している。このシナリオでは、二人の話者Ａ，Ｂが約４０秒間に亘って種々のアクションを行なう。前記シナリオは、具体的には以下の通りである。

時刻ｔ１：Ａ氏がロボット１０の視野内に入る。視覚モジュール３０がＡ氏の顔を検出して視覚イベントを生成

し、アソシエーションモジュール６０により視覚ストリームが生成される。
時刻ｔ２：Ａ氏がロボット１０に対して話し始める。聴覚モジュール２０がＡ氏の声を検出して聴覚イベントを生成し、アソシエーションモジュール６０により聴覚ストリーム６５が生成され、さらにアソシエーションストリーム６７が生成される。これにより、ロボットの視聴覚サーボが行なわれる。
時刻ｔ３：Ｂ氏がロボット１０の視野外で話し始める。聴覚モジュール２０が（見えない）Ｂ氏の声を検出して、聴覚イベントを生成し、アソシエーションモジュール６０により聴覚ストリームが生成される。
時刻ｔ４：Ａ氏が移動して、物陰に隠れる。これにより、視覚モジュール３０がＡ氏の視覚イベントを生成しなくなり、Ａ氏の視覚ストリームが途切れるが、アソシエーションストリームは所定時間の間存続する。
時刻ｔ５：Ａ氏が再び物陰から現われる。これにより、視覚モジュール３０がＡ氏の視覚イベントを再び生成し、アソシエーションモジュール６０により、再びアソシエーションストリーム６７が生成される。その後、Ａ氏は話を止めて、再び物陰に隠れる。視覚モジュール３０がＡ氏の視覚イベントを生成しなくなり、Ａ氏の視覚ストリームが途切れるので、所定時間後にアソシエーションが解除され、アソシエーションストリーム６７が消滅する。
時刻ｔ７：聴覚ストリームをトリガーとして、ロボット１０が話をしているＢ氏の方を向く。
時刻ｔ８：ロボット１０がＢ氏を視野内に捉える。視覚モジュール３０がＢ氏の視覚イベントを生成し、アソシエーションモジュール６０によりＢ氏の視覚ストリームが生成され、さらにＢ氏のアソシエーションストリーム６７が生成される。
時刻ｔ９：Ａ氏が話をしながら、ロボット１０の視野内に入ってくる。聴覚モジュール２０及び視覚モジュール３０がＡ氏の聴覚イベント及び視覚イベントを生成し、アソシエーションモジュール６０がＡ氏の聴覚ストリーム及び視覚ストリームが生成される。
時刻ｔ１０：Ｂ氏が話を止める。聴覚モジュール２０がＢ氏の聴覚イベントを生成しなくなり、アソシエーションモジュール６０がＢ氏のアソシエーションを解除してＢ氏の聴覚ストリームは消滅し、視覚ストリームのみが残る。そして、ロボット１０がアテンションをＡ氏に向けると共に、同様にしてＡ氏のアソシエーションストリーム６７が生成される。
【００６２】このようにして、上述したシナリオにおいては、本発明によるロボット視聴覚システムにおいては、以下のような特徴を有することが分かる。
１．　時刻ｔ１及びｔ６にて、新しいアソシエーションストリームが生成されると、アテンション制御モジュール６４におけるアテンションが新しいアソシエーションに向けられる。
２．　時刻ｔ４，ｔ５にて、Ａ氏が見えなくなることにより、アソシエーションストリームの視覚ストリームが欠落したときであっても、アソシエーションが存続していることにより、聴覚ストリームによるＡ氏のトラッキングが継続され得る。
３．　時刻ｔ６，ｔ１１にて、アソシエーションストリームが消滅することにより、アソシエーションストリームの次に優先度の高い聴覚ストリームによりアテンション制御が行なわれ、話者のトラッキングが行なわれ、図１３に示すように、ロボット１０がトラッキングの対象である話者に正対して、当該話者からの音声をマイク１５の正面方向により確実に検出することができるようになっている。
４．　時刻ｔ９以降、二人の話者Ａ氏及びＢ氏は、同時にカメラ１５の視野内に収まる程度（方向差約２０度）に接近しているが、この場合でも、二人の聴覚ストリーム，視覚ストリーム及びアソシエーションストリームは、それぞれ明確に別個に生成され、各話者のトラッキングが確実に行なわれる。
【００６３】このようにして、人型ロボット１０は、聴覚イベント２８及び視覚イベント３９が生成される場合には、これらを互いに関連付けて、アソシエーションストリーム６７を生成して、このアソシエーションストリーム６７に基づいてアテンション制御を行なうことにより、ロボットの視聴覚サーボを行なうことになる。従って、従来の聴覚サーボまたは視覚サーボの場合と比較して、聴覚及び視覚の双方を使用することによって、話者をより正確に追跡することが可能になる。また、途中で話者が物陰に隠れたり視野外に移動して見えなくなって、図１４（Ｂ）（視覚イベントの第一候補のみを示す）に示すように視覚ストリームによるトラッキングが途切れた場合には、図１４（Ｃ）に示すように、聴覚ストリームによるアソシエーションストリーム６７によって、当該話者を確実にトラッキングすることかできるので、常に複数の話者を聴覚及び視覚により認識していると共に、複数の話者のうちの一人の話者を追跡したり、あるいは途中で他の話者に切り換えて追跡することができる。
【００６４】なお、図１４（Ｂ）において、時刻ｔ４及びｔ５の間では視覚ストリームが途切れ、また時刻ｔ６からｔ７までの間は話者がロボット１０の視野外に居ることから、視覚ストリームに基づいて、話者のトラッキングを行なうことはできないが、図１４（Ｃ）に示す聴覚ストリームを参照することによって、話者のトラッキングを確実に行なうことができる。また、図１４（Ｃ）において、時刻ｔ３が２３秒付近まで、そして３４秒付近から時刻ｔ１０の間は、正しくＡ氏及びＢ氏の二本の聴覚ストリームが分離されているが、時刻ｔ８及びｔ６の周辺では、誤った聴覚ストリームが生成されている。

また、時刻ｔ５から１７秒付近までの間は、Ａ氏の移動及びロボット１１の水平回転が同時に行なわれているため、話者の移動及びモータノイズそしてそのエコーにより音源からの音響信号の品質が低下しており、二人の話者の定位はあまり正確ではない。このような場合でも、図１４（Ｂ）に示す視覚ストリームを参照することにより、話者のトラッキングを確実に行なうことができる。このようにして、聴覚ストリーム及び視覚ストリームが互いに関連付けられてアソシエーションストリームが生成される場合には、聴覚及び視覚の双方を使用して、ロボットの視聴覚サーボを行なうことによって、聴覚ストリーム及び視覚ストリームがそれぞれ有する曖昧性が互いに補完されることにより、所謂ロバスト性が向上し、複数の話者であっても、各話者をそれぞれ確実に知覚して、トラッキングを行なうことができる。

【００６５】また、コンパニオン用ロボットとしての人型ロボット１０は、パーティ受付ロボットと顔データベース３８を共用し、あるいはパーティ受付ロボットの顔データベース３８が転送または複写されるようにしてもよい。この場合、コンパニオン用ロボットとしての人型ロボット１０は、パーティ参加者全員を常に顔識別によって認識することができる。

【００６６】このようにして、本発明実施形態による人型ロボット１０によれば、聴覚モジュール２０及び視覚モジュール３０からの聴覚イベント及び視覚イベントに基づいて、アソシエーションモジュール６０が聴覚ストリーム，視覚ストリームそしてアソシエーションストリームを生成することによって、複数の対象である話者を視聴覚により認識しているので、聴覚または視覚のいずれか一方のみによるサーボの場合と比較して、より正確に話者の追跡を行なうことができると共に、何れかのイベントが欠落したり明確に認識できなくなった場合には、例えば話者が移動して「見えなく」なった場合でも聴覚により、また話者が話をせず「聞こえなく」なった場合でも視覚により、リアルタイムに複数の話者を聴覚的及び／又は視覚的にトラッキングすることができる。

【００６７】上述した実施形態において、人型ロボット１０は、４ＤＯＦ（自由度）を有するように構成されているが、これに限らず任意の動作を行なうように構成されたロボットに本発明によるロボット聴覚システムを組み込むことも可能である。また、上述した実施形態においては、本発明によるロボット視聴覚システムを人型ロボット１０に組み込んだ場合について説明したが、これに限らず、犬型等の各種動物型ロボットや、その他の形式のロボットに組み込むことも可能であることは明らかである。

【００６８】

【発明の効果】以上述べたように、この発明によれば、聴覚モジュール，視覚モジュール及びモータ制御モジュールと、アソシエーションモジュール及びアテンション制御モジュールとの連携によって、聴覚及び視覚の双方を使用して、ロボットの視聴覚サーボを行なうことにより、ロボットの聴覚及び視覚がそれぞれ有する曖昧性が互いに補完されることになり、所謂ロバスト性が向上し、複数の話者であっても各話者をそれぞれ知覚することができる。また、例えば聴覚イベントまたは視覚イベントの何れか一方が欠落したときであっても、視覚イベントまたは聴覚イベントのみに基づいて、対象である話者をアソシエーションモジュールが知覚することができるので、リアルタイムにモータ制御モジュールの制御を行なうことができる。さらに、聴覚ストリーム及び視覚ストリームのうち、状況に応じて、双方または一方のみを利用して、話者のトラッキングを行なうことにより、常により一層正確な話者のトラッキングを行なうことができると共に、同時に複数の聴覚ストリーム及び視覚ストリームが存在していても、これらの聴覚ストリーム及び視覚ストリームに基づいて、そのときの状況に応じて、これらの聴覚ストリーム及び視覚ストリームの何れかを適宜に利用することにより、各話者のトラッキングをより確実に行なうことができる。これにより、本発明によれば、対象に対する視覚及び聴覚の追跡を行なって、視覚及び聴覚の双方を使用してロボットの視聴覚サーボを行なうようにした、極めて優れたロボット視聴覚システムが提供される。

【図面の簡単な説明】

【図１】この発明によるロボット聴覚装置の第一の実施形態を組み込んだ人型ロボットの外観を示す正面図である。

【図２】図１の人型ロボットの側面図である。

【図３】図１の人型ロボットにおける頭部の構成を示す概略拡大図である。

【図４】図１の人型ロボットにおけるロボット視聴覚システムの電気的構成を示すブロック図である。

【図５】図４におけるブロック１の聴覚モジュールを拡大して示す電気的構成のブロック図である。

【図６】図４におけるブロック２の視覚モジュールを拡大して示す電気的構成のブロック図である。

【図７】図４におけるブロック３のモータ制御モジュールを拡大して示す電気的構成のブロック図である。

【図８】図４におけるブロック４の対話モジュールを拡大して示す電気的構成のブロック図である。

【図９】図４におけるブロック５のアソシエーションモジュールを拡大して示す電気的構成のブロック図である。

【図１０】図４のロボット視聴覚システムにおけるパーティ受付ロボットとしての動作例を示す図である。

【図１１】図４のロボット視聴覚システムにおける（Ａ）聴覚モジュール，（Ｂ）視覚モジュールのビューアの画面例を示す図である。

【図１２】図４のロボット視聴覚システムにおける

（Ａ）モータ制御モジュール，（Ｂ）アソシエーションモジュールのビューアの画面例を示す図である。

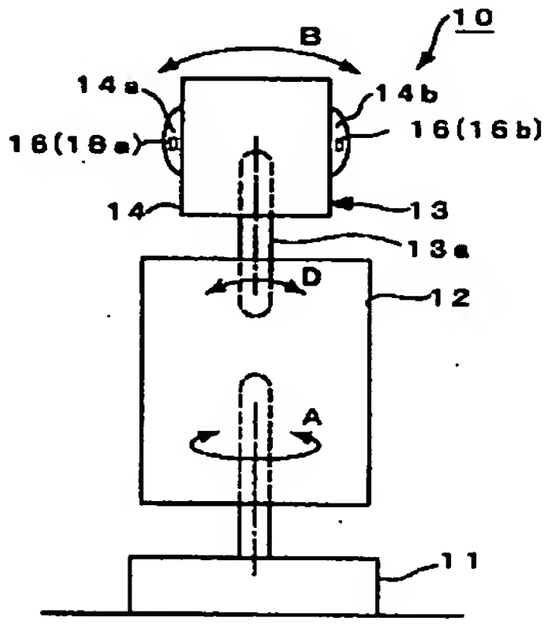
【図１３】図４のロボット視聴覚システムにおけるコンパニオン用ロボットとしての動作例を示す各時刻における（Ａ）レーダチャート，（Ｂ）ストリームチャート及び（Ｃ）カメラ画像を示す図である。

【図１４】図１３の動作例における（Ａ）ロボット方向，（Ｂ）視覚ストリーム及び（Ｃ）聴覚ストリームを示すグラフである。

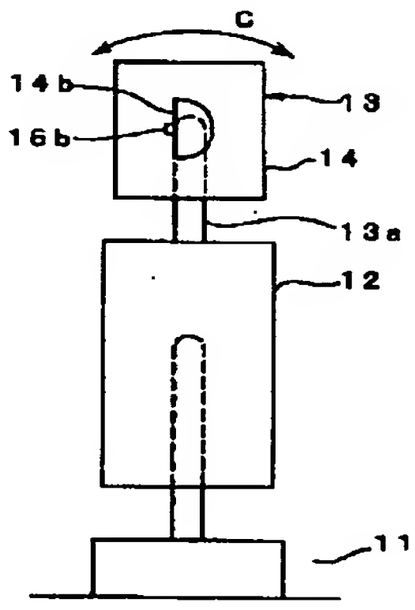
【符号の説明】

１０　人型ロボット
１１　ベース
１２　胴体部
＊１３　頭部
１３ａ　連結部材
１４　外装
１５　カメラ（ロボット視覚）
１６，１６ａ，１６ｂ　マイク（ロボット聴覚）
１７　ロボット視聴覚システム
２０　聴覚モジュール
３０　視覚モジュール
４０　モータ制御モジュール
５０　対話モジュール
６０　アソシエーションモジュール
６４　アテンション制御モジュール
＊７０　ネットワーク

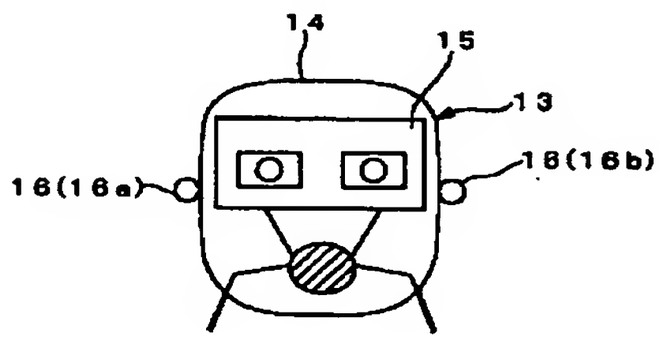
【図１】



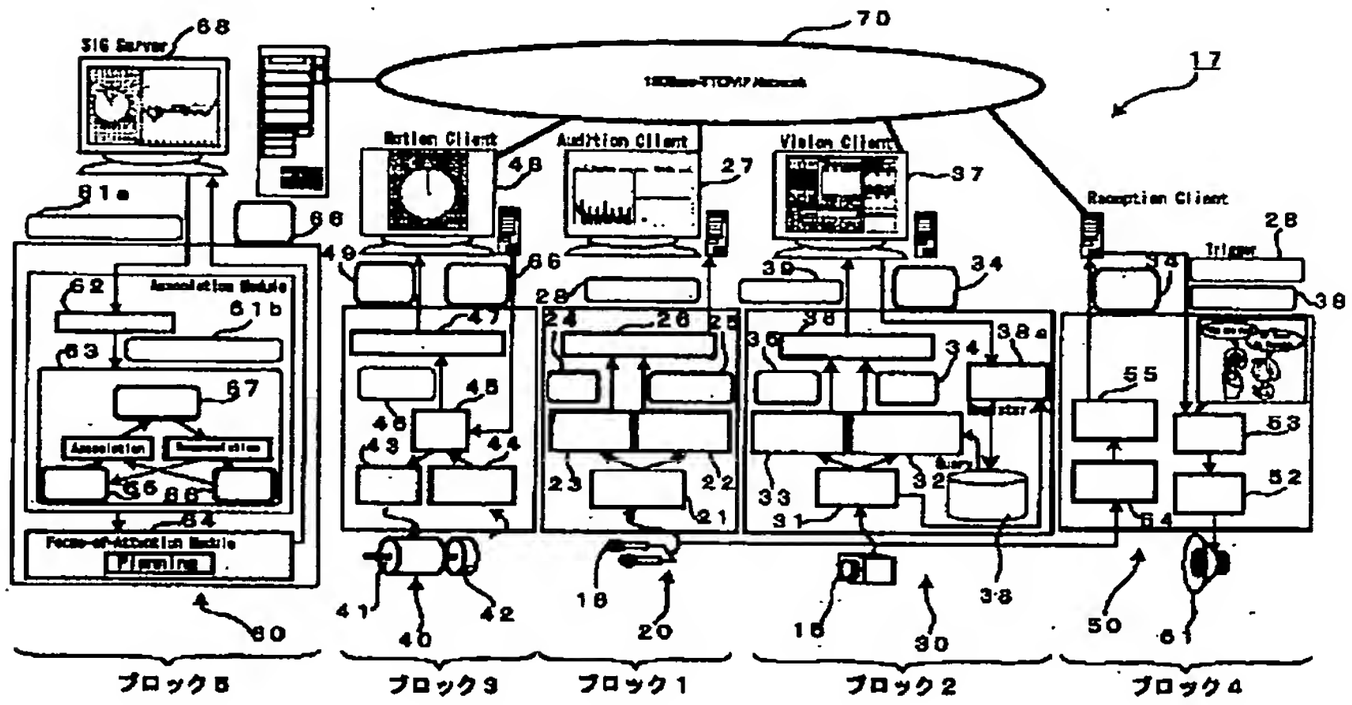
【図２】



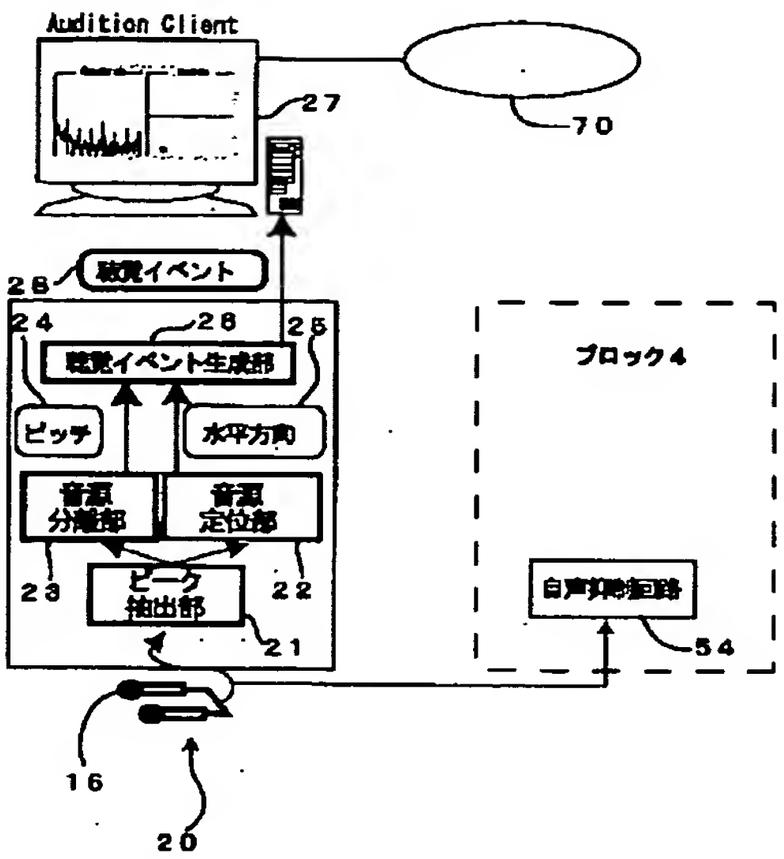
【図３】



【図４】

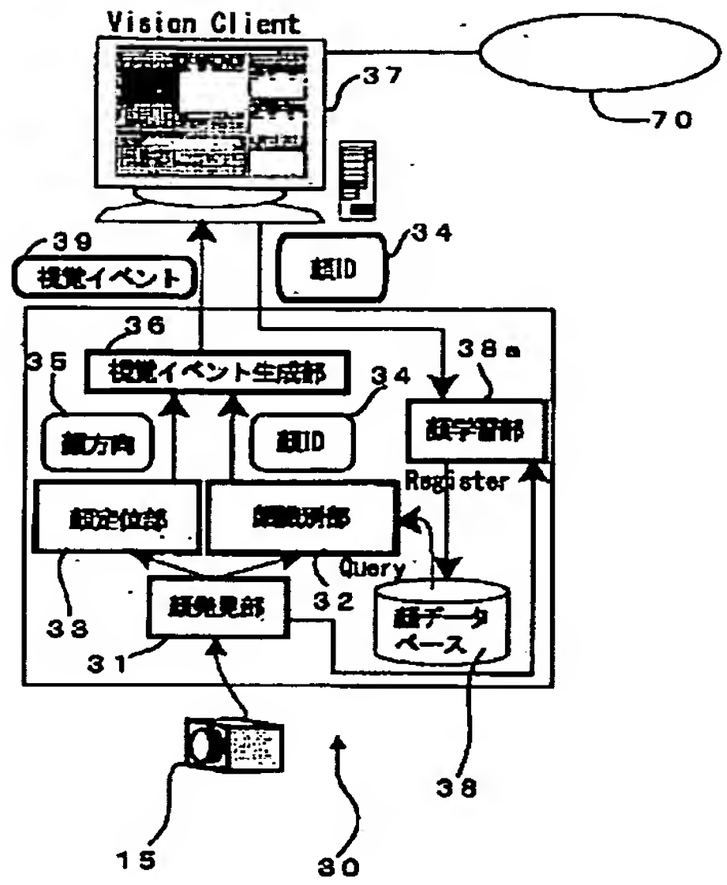


【図５】



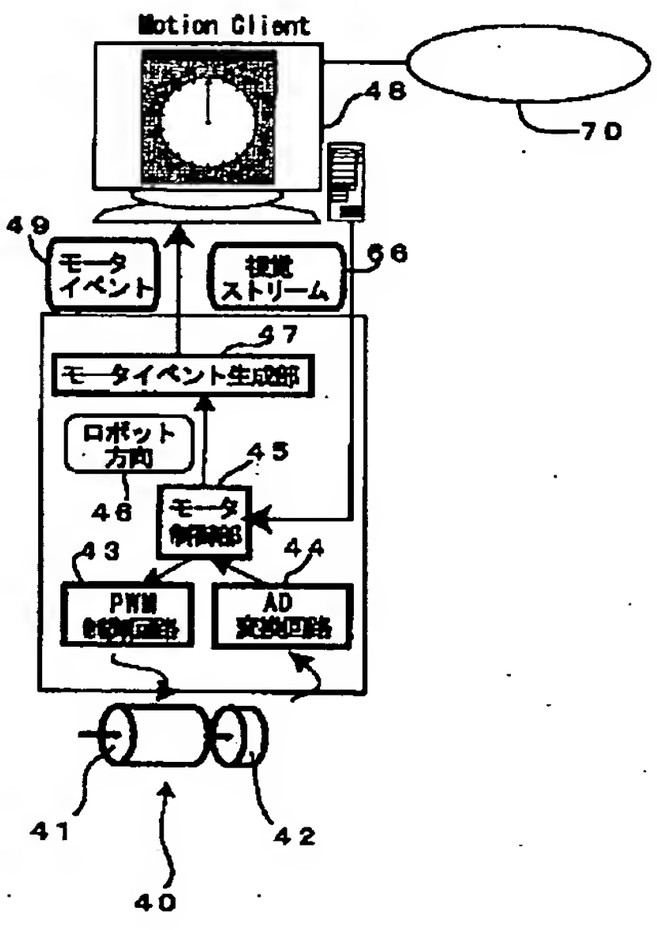
ブロック１

【図６】



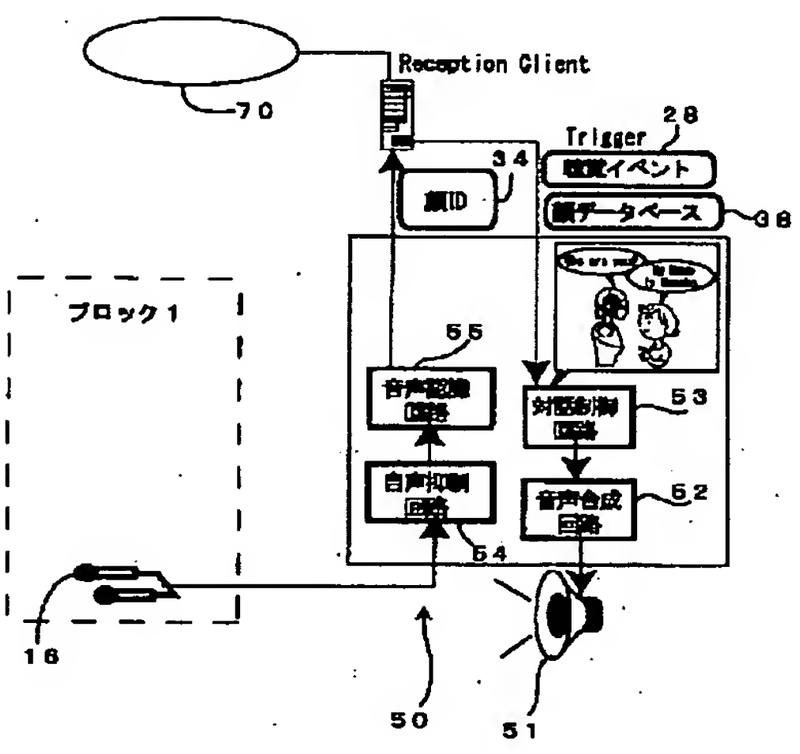
ブロック２

【図７】



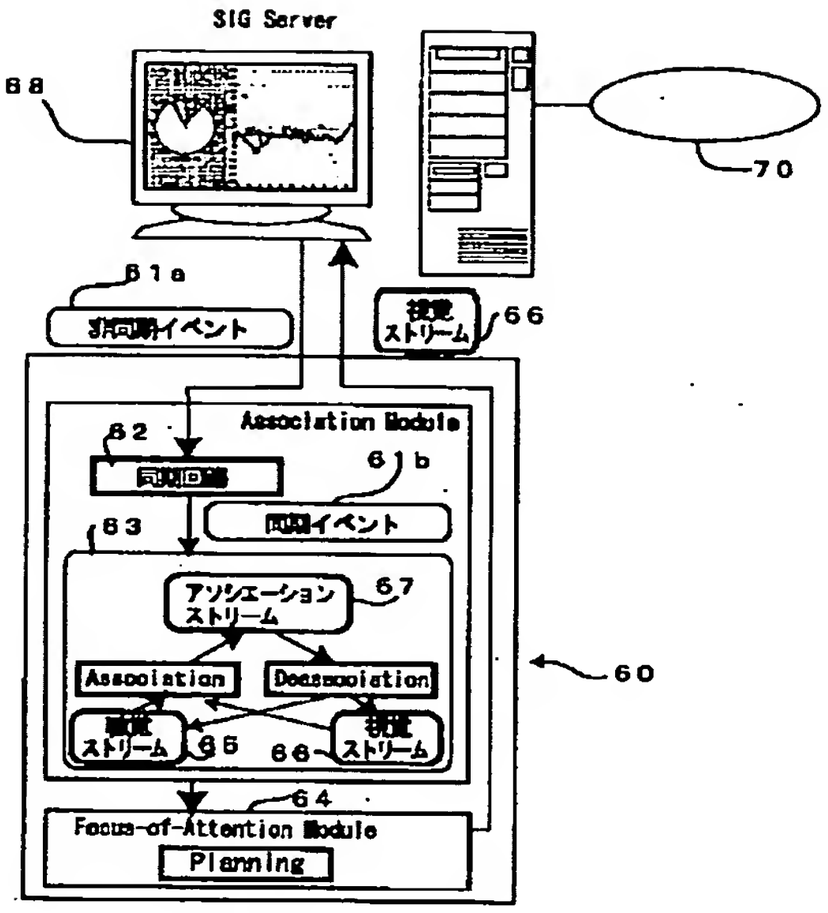
ブロック３

【図８】



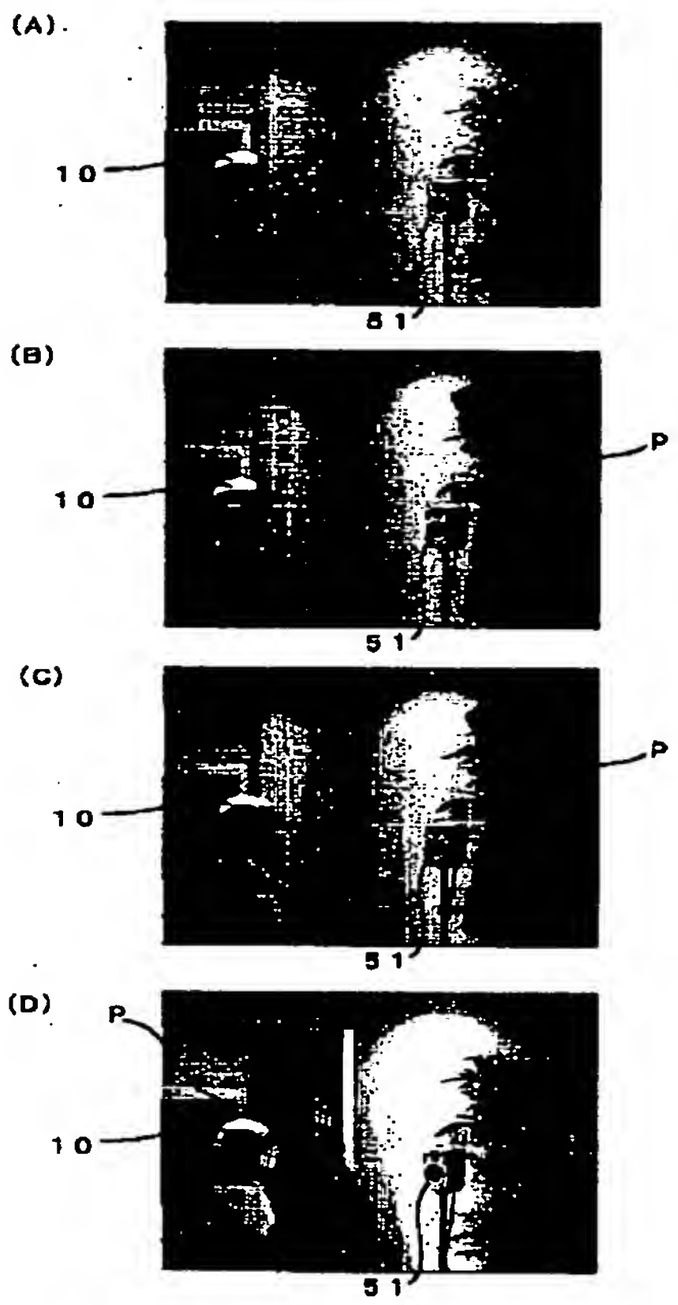
ブロック４

【図９】

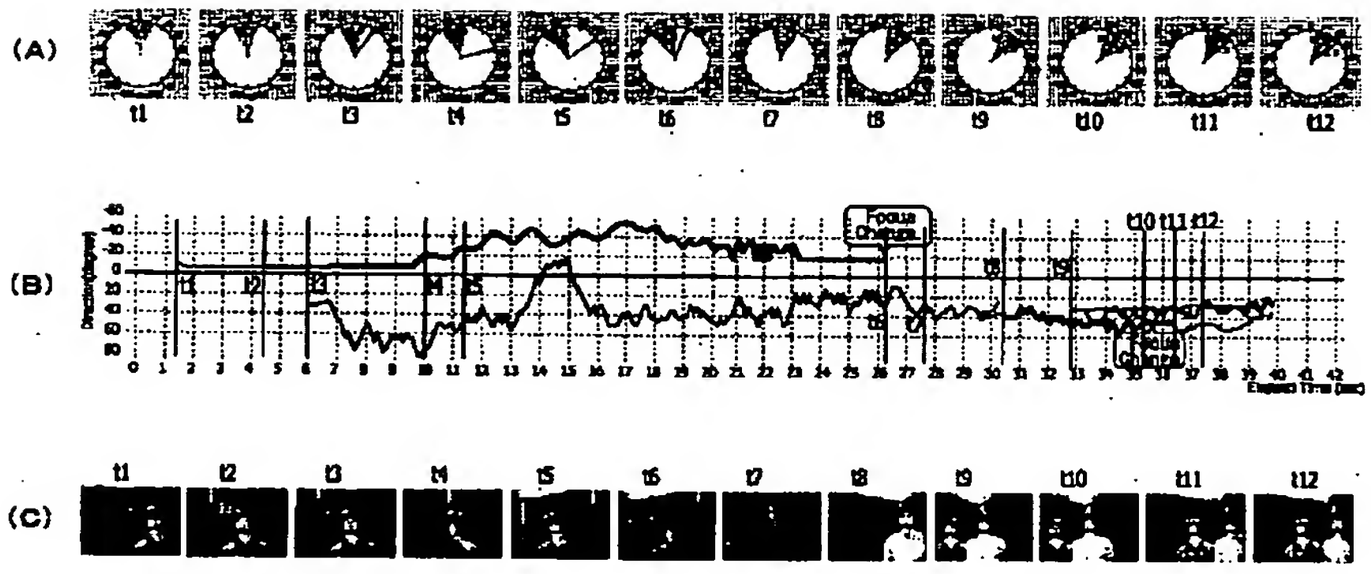


ブロック５

【図１０】

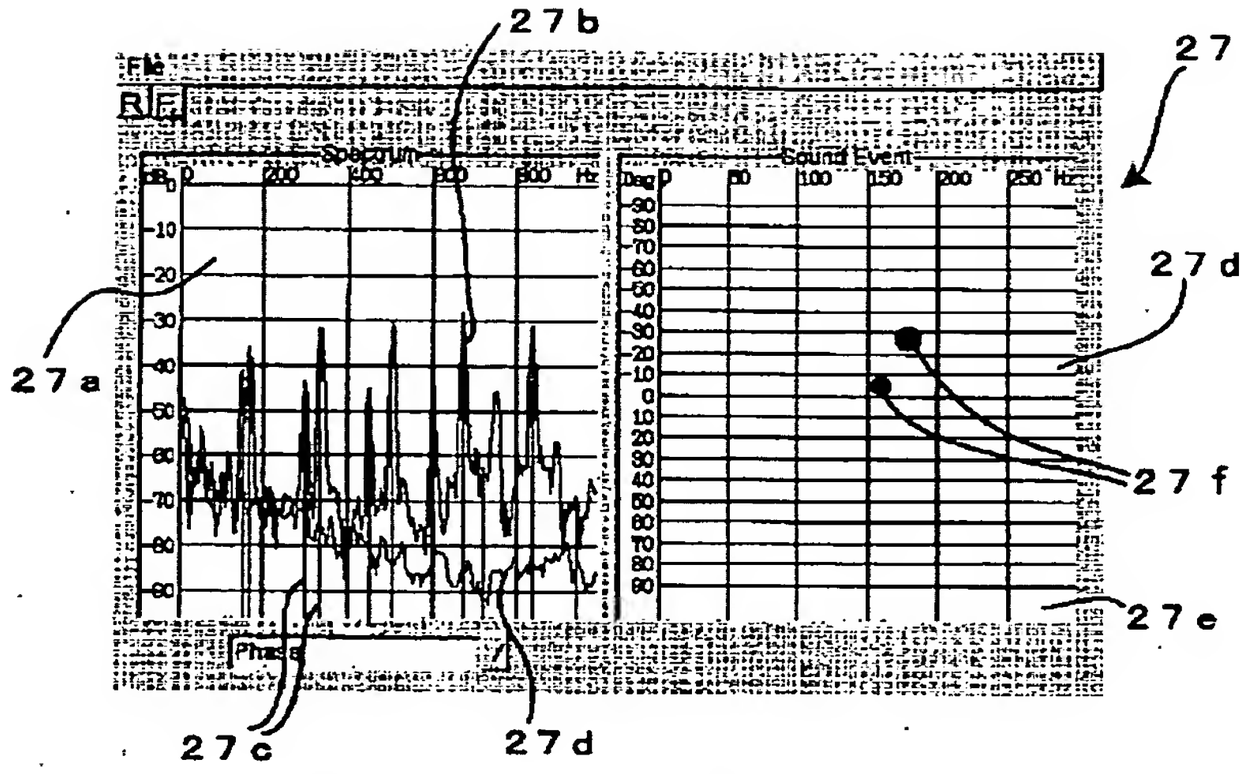


【図１３】

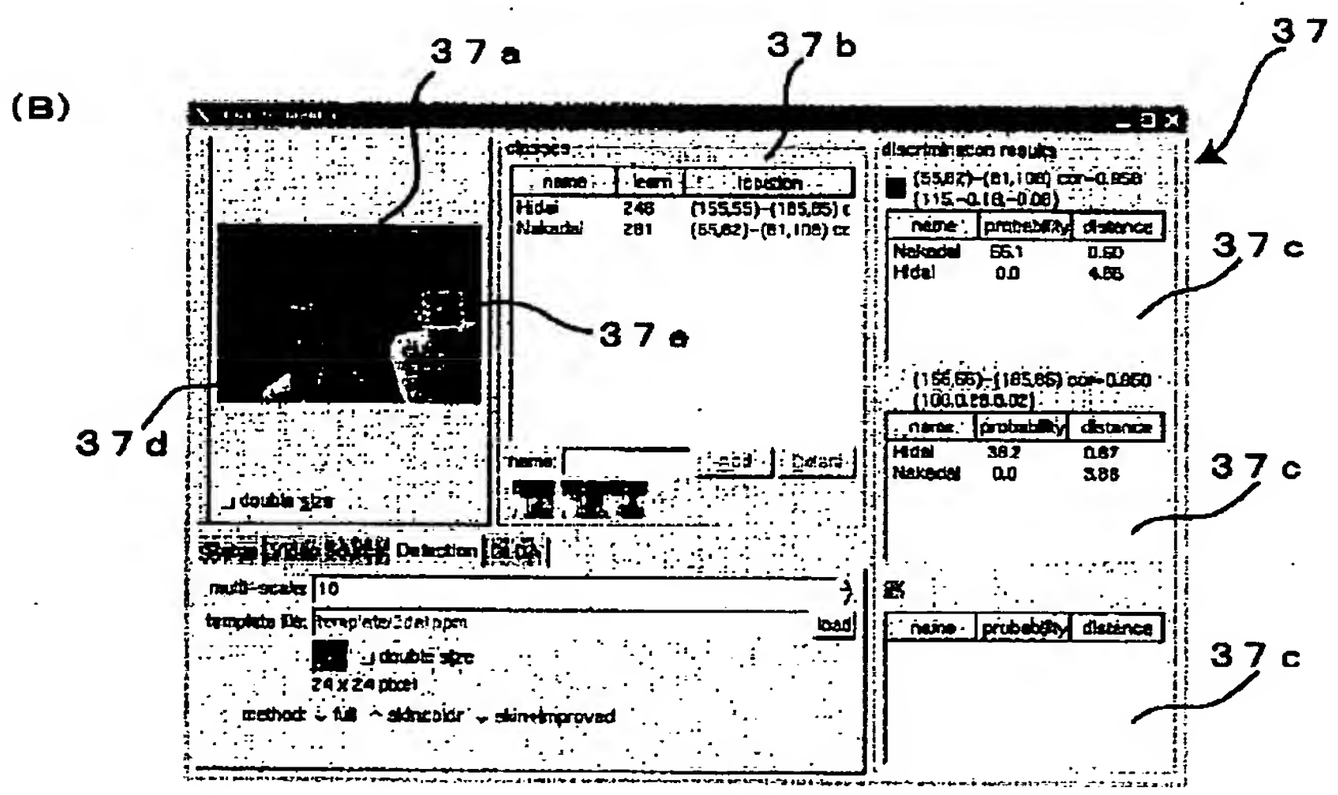


【図１１】

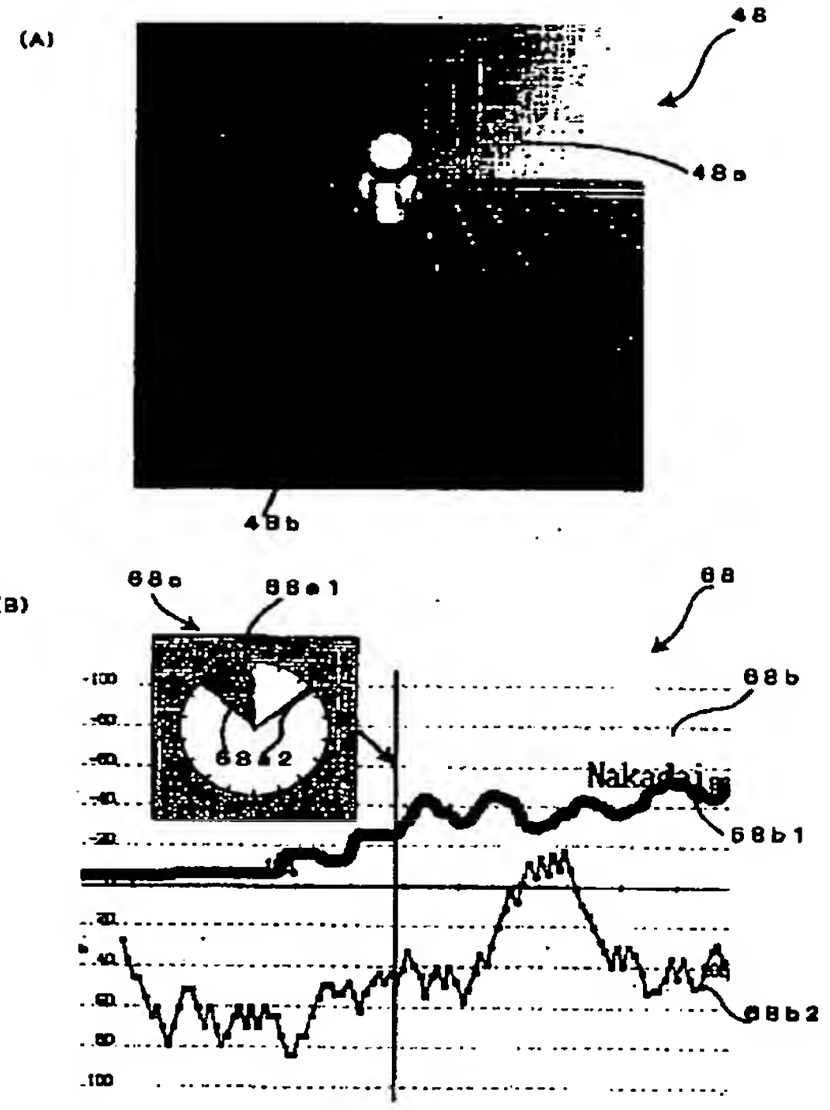
(A)


27b
27
27a
27d
27f
27e
27c
27d
Spectrum
Sound Event
Phase

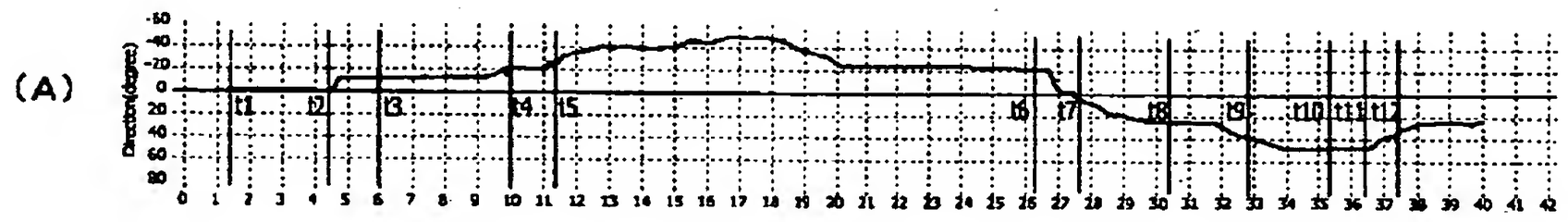
(B)

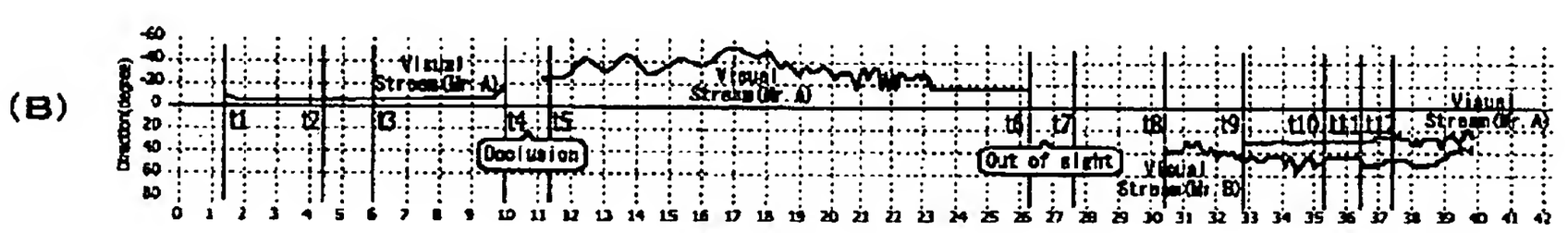

37a
37b
37
37e
37d
37c
37c
37c
classes
discrimination results
name
learn
location
probability
distance
Hidai
Nakadai
double size
load
Detect
multi-scale
template file
method

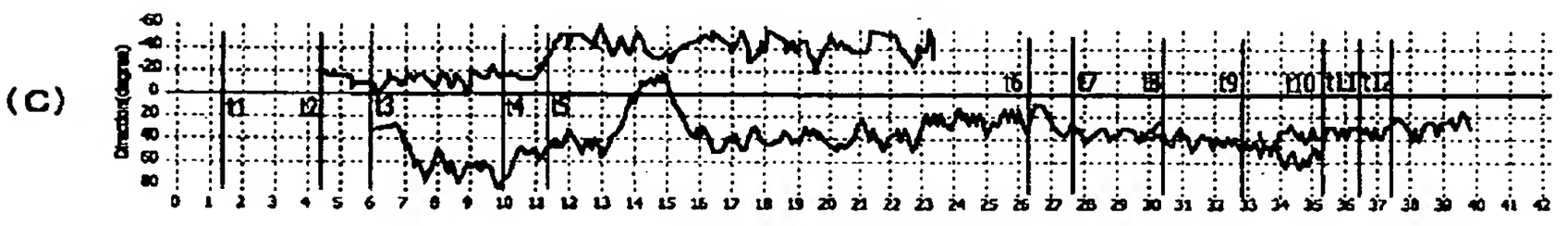
【図１２】


(A)
48
48b
48b
(B)
68c
68c1
68
68b
68c2
68b1
68b2

【図１４】


(A)
Direction(degree)
t1
t2
t3
t4
t5
t6
t7
t8
t9
t10
t11
t12


(B)
Direction(degree)
Visual Stream(Gr. A)
Occlusion
Out of sight
Visual Stream(Gr. B)
Visual Stream(Gr. A)


(C)
Direction(degree)

フロントページの続き

(51)Int.Cl.'		識別記号	FI			テーマコード（参考）
Ｇ０６Ｔ	7/60	１５０	Ｇ０６Ｔ	7/60	１５０Ｂ	５Ｌ０９６
Ｇ１０Ｌ	11/04		Ｈ０４Ｎ	7/18	Ｚ	
	13/00		Ｇ１０Ｌ	3/00	Ｃ	
	15/28				Ｑ	
	17/00				５１１	
	15/00				５４５Ｆ	
	15/22				５５１Ｈ	
	15/20				５７１Ｔ	
	21/02			3/02	３０１Ｃ	
	15/02			9/00	３０１Ａ	
Ｈ０４Ｎ	7/18					

Ｆターム(参考)　3C007 AS34 AS36 CS08 JS03 KS04
KS08 KS18 KS20 KS39 KT01
KT11 KT15 LT08 NS01 WA02
WA03 WB19 WC07 WC16
5B057 AA05 BA02 CA12 CA16 DA06
DB02
5C054 AA01 CA04 CA08 CC05 CD03
CG06 EF06 FC12 FF07 HA04
5D015 AA03 CC13 DD02 EE04 KK01
KK04 LL06
5D045 AB11
5L096 BA05 CA02 FA69 HA09